Discrete Applied Mathematics 37/38 (1992) 359-385
North-Holland

Cost-performance tradeoffs for interconnection networks*

Clyde P. Kruskal

Department of Computer Science, University of Maryland, College Park, MD 20742, USA

Marc Snir

IBM Research Division, T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, USA

Received 21 June 1989
Revised 25 February 1991

Abstract

Kruskal, C.P. and M. Snir, Cost-performance tradeoffs for interconnection networks, Discrete Applied Mathematics 37/38 (1992) 359-385.

A major component of a large-scale parallel computer is the interconnection network that connects processors to memories in a shared-memory machine, or processors to processors in a multicomputer. This paper formally studies the relationship between network topology and network performance. Rectangular banyan networks are shown to provide maximum bandwidth/cost ratio for symmetric traffic. For their cost, contracting banyan networks are shown to provide maximum bandwidth up to a constant factor for semisymmetric traffic. For a restricted class of networks, contracting banyan networks are shown to provide exactly maximum bandwidth for semisymmetric traffic. Rectangular banyan networks are shown to provide optimal delay-to-cost tradeoffs for symmetric traffic. It is shown that, in many situations, optimal bandwidth is achieved by using a unique path to route information between each input-output pair.

1. Introduction

A major component of a large-scale parallel computer is the *interconnection network* that connects processors to memories in a shared-memory machine, or processors to processors in a multicomputer. Such a network often consists of switches interconnected by links. Both the cost and performance of a network are

* Preliminary versions of some of the material in this paper have appeared in the *11th Annual International Symposium on Computer Architecture* (1984), *Current Advances in Distributed Computing and Communications* (1987), and the *1st Annual Symposium on Parallel Algorithms and Architectures* (1989).

affected by its topology. A huge variety of network topologies have been proposed in the literature, and studied in an ad hoc manner. A systematic, comparative study of these proposals is hampered by the lack of good performance criteria. Various topological parameters, such as diameter or cutwidth, have been used as "measures of goodness" for network topologies. However, it is clear that network performance depends on the type of traffic it supports; for example, a star topology that performs well for centralized traffic will have poor performance for uniformly distributed traffic.

We propose in this paper a formal approach to the study of the relationship between network topology and network performance. We characterize the traffic pattern in terms of the relative frequency of communications between each pair of nodes. We assume that networks are "pin-limited"; there is a fixed upper bound on the number of links incident to any one node or switch. Network *bandwidth* and *delay* are taken to be the main figures of merit. We further assume that the *cost* of a network is essentially proportional to its number of links.

Formal definitions are given in Section 2. Bandwidth is defined to depend only on the network topology and the traffic pattern. Such a definition obviously ignores many aspects of network design—for example, the network control mechanism. However, we validate this definition in Appendix A by showing that such bandwidth can be achieved by reasonable control mechanisms, in various settings.

The definition of bandwidth can be used to analyze the performance of specific networks, under given traffic patterns. More importantly, the definition can be used to approach the following "topology optimization" problem: Given a traffic pattern, and given a bound on network cost, find the topologies that achieve maximal bandwidth. A first step is to obtain the value of an optimal solution: Given a traffic pattern, and given a bound on network costs, compute the maximal achievable bandwidth.

It turns out that traffic patterns can be usefully characterized in terms of their *entropy*. In Section 3 a basic inequality is derived that relates bandwidth to cost and traffic entropy. This relation is shown in Section 4 to be precisely tight for the important particular case of *symmetric traffic*, where each input is equally likely to communicate with each output, and vice versa. The optimal networks are characterized as a family of *rectangular banyan networks*. More general traffic distributions are considered in Section 5. The basic inequality is shown to be tight, up to a constant factor, in the particular case where traffic from each input is equally distributed over all outputs (however, distinct inputs may generate distinct amounts of traffic): a construction is given for networks that are optimal (up to a constant factor). The construction is extended to arbitrary traffic patterns; however, it is not optimal in the general case. One reason for the gap is that traffic entropy is not a sufficient characterization: We exhibit two traffic patterns with the same entropy but different cost-to-bandwidth tradeoffs.

In Section 6 we examine another performance measure, namely delay. Definitions are given for delay in terms of the network topology and the traffic distribution.

These definitions are motivated by a suitable stochastic model. Basic inequalities are derived for delays. Rectangular banyan networks are shown to achieve optimal delay-to-cost tradeoffs for symmetric traffic.

One of the tools used to derive the optimality results is a characterization of the bandwidth optimization problem as a multicommodity flow problem. This enables us to derive an integral solution theorem that implies that, in many situations, an optimal bandwidth is achieved by using a unique path to route traffic between each input-output pair. This derivation is presented in Section 7.

2. Definitions

We shall consider in this paper message-switched communication networks. Messages are generated by their input, and routed to their output via a path of store-and-forward nodes (switches) that are connected by unidirectional communication lines. We assume that the set of inputs is disjoint from the set of outputs—the inputs may be processors and the outputs may be memory modules in a shared-memory multiprocessor. We represent an *interconnection network* as a directed graph: Nodes in the graph correspond to switches and links correspond to communication lines. There is a set $\{1,\ldots,M\}$ of M *input nodes* with indegree zero and a set $\{1,\ldots,N\}$ of N *output nodes* with outdegree zero. The *cost* $C=C(G)$ of a network G is defined to be the number of links in G. We denote by $I(u)$ the set of links incoming node u, and by $O(u)$ the set of links outgoing node u. Thus $|I(u)|$ ($|O(u)|$) is the indegree (outdegree) of node u, and

$$C=\sum_u |I(u)| = \sum_u |O(u)|.$$

We characterize a communication pattern in terms of a *traffic distribution* function π; $\pi_{i,j}$ is the relative frequency of traffic from input i to output j. We have $\sum_i \sum_j \pi_{i,j}=1$. In a *symmetric* traffic distribution $\pi_{i,j}=1/MN$ for each i, j.

We do not make, at this point, any assumption about the mechanism whereby messages are generated and routed. However, for the sake of concreteness, one can think of the following two models:

- The *discrete* model. A large number K of messages are generated at the inputs; input i has $K\pi_{i,j}$ messages to send to output j. A schedule is computed offline to route these messages through the network. The schedule specifies which message is routed on each link at each cycle. We are interested in the ratio between K and the time needed to transfer the K messages, for large K.
- The *continuous* model. Messages are continuously generated by inputs by a stochastic process; $\pi_{i,j}$ is the probability that a message is generated at input i for output j. A route is probabilistically assigned to the message, and the message is forwarded (obliviously) on that route. A simple service policy (e.g. FCFS) is used at each switch. We are interested in least upper bounds on the traffic intensity (the average number of new messages entering the network per time unit) in steady-state.

Let $\Pi_{i,j}$ be the set of directed paths connecting input i to output j. A routing algorithm associates with each message sent from input i to output j a path $p \in \Pi_{i,j}$. We characterize the routing algorithm in terms of a *route distribution* ϱ; $\varrho(p)$ is the relative frequency of traffic using path p. We can assume without loss of generality that only simple paths are used for routing; deleting loops from paths can only improve performance. In the deterministic model, the product $K\varrho(p)$ is the number of messages using path p; in the probabilistic model $\varrho(p)$ is the probability that a message is routed via path p. We have

$$\sum_{p \in \Pi_{i,j}} \varrho(p) = \pi_{i,j}. \tag{1}$$

We say that a route distribution ϱ is *consistent* with traffic distribution π, if it fulfills equation (1).

The route distribution determines the load on each node and each link in the network. We define the *relative load* $\omega_\varrho(e)$ of link e to be the ratio between the number of messages forwarded on link e and the total number of messages processed by the network:

$$\omega_\varrho(e) = \sum_{e \in p} \varrho(p)$$

(the sum is taken over all paths using edge e). In the discrete model, link e will forward a total of $K\omega_\varrho(e)$ messages. In the continuous model, $\omega_\varrho(e)$ is the probability that a message uses link e.

We assume that there is a fixed bound on the degree of nodes in a network—this corresponds to actual pin count constraints of components. Although this constraint will apply to both the *indegree* and the *outdegree*, it will usually be sufficient to bound either one in order to obtain our lower bound results.

We also assume that links are the main limiting factor on performance. Therefore, we will assume that a switch can simultaneously forward up to one message (or up to one message on the average) per time unit, on each of its outgoing links. Internally, a switch is assumed to have unlimited buffering capacity.

Assume that τ new messages enter the network on the average per time unit. Then the average number of messages forwarded on link e each time unit is $\tau\omega_\varrho(e)$. This implies that $\tau\omega_\varrho(e) \leq 1$. Thus, $\tau \leq \min_e 1/\omega_\varrho(e)$. Accordingly, we define the *bandwidth* B_ϱ of a network G for a route distribution ϱ to be equal to

$$B_\varrho = \min_e \frac{1}{\omega_\varrho(e)}.$$

The bandwidth $B = B(\pi)$ for a traffic distribution π is taken to be

$$B = \max_\varrho B_\varrho,$$

where the maximum is taken over all route distributions ϱ that are consistent with π. The bandwidth B is a function of the network topology and of the distribution

π. In the continuous model, B is an upper bound on the average number of new messages that enter the network per time unit (in steady-state). In the discrete model, K/B is a lower bound on the time needed to transfer K messages.

In order to justify the proposed definition of bandwidth we would like to argue that the bandwidth B is not merely an upper bound on the traffic intensity that the network can support for a given traffic distribution, but is the least such upper bound. This is demonstrated in Appendix A.

3. Entropy and basic inequalities

We derive in this section lower bounds on network cost, as a function of traffic distribution and bandwidth. The lower bounds are mostly in terms of the entropy of the traffic distribution.

We define the *mean path length* of the network to be the average number of links traversed by messages in the network (assuming a given route distribution ϱ); this is equal to

$$d_\varrho = \sum \varrho(p)\,|p|.$$

Lemma 3.1. *Let G be a network of cost C. Then, for any route distribution ϱ*

$$C \geq d_\varrho \cdot B_\varrho.$$

Equality holds if and only if all links have equal relative load $\omega_\varrho(e) = 1/B_\varrho$.

Proof. We have

$$d_\varrho = \sum_p \varrho(p)\,|p| = \sum_e \sum_{e \in p} \varrho(p)$$

$$= \sum_e \omega_\varrho(e) \geq \sum_e \frac{1}{B_\varrho} = \frac{C}{B_\varrho}. \qquad \square$$

The last inequality reflects an obvious relation: the traffic intensity is bounded by the total number of links divided by the average number of links traversed by a message.

We assume that there is a fixed upper bound on the degree of nodes in a network—this corresponds to actual pin count constraints of components. This constraint applies both to indegree and outdegree. However, it is sufficient to bound either indegree or outdegree in order to obtain the lower bounds of this section. We assume in this section that G is a network with M inputs, N outputs, and outdegree $\leq k$. All results apply dually to networks with N inputs, M outputs, and indegree $\leq k$. In this and following sections, k is assumed to be fixed when "O" notation is used.

There is an obvious constraint on the bandwidth of such networks: There are at

most kM links connected to the input nodes, so that the bandwidth is at most kM. This is formally stated in the following lemma:

Lemma 3.2. *Let G be a network with M inputs and outdegree bounded by k. Then*

$$B \leq kM.$$

We recall the following definitions and results from coding theory. Let $p_1, \ldots, p_n$ be a probability distribution. The *entropy* of this distribution is

$$H_k(p_i) = -\sum_{i=1}^{n} p_i \log_k p_i.$$

We have

$$0 \leq H_k(p_i) \leq \log_k n.$$

Equality obtains on the left-hand side iff the distribution is degenerate, i.e., $p_i \in \{0,1\}$. Equality obtains on the right-hand side iff $p_i = 1/n$, $i = 1, \ldots, n$.

Let $p_{i,j}$ be a probability distribution, and let $p_i = \sum_j p_{i,j}$ be the *marginal distribution*. We denote by

$$H_k(p_{r,j} \mid r = i) = -\sum_j \frac{p_{i,j}}{p_i} \log_k \frac{p_{i,j}}{p_i}$$

the entropy of the *conditional distribution* of $p_{i,j}$, for fixed i. The *conditional entropy* of $p_{i,j}$, given p_i, is defined as

$$H_k(p_{i,j} \mid p_i) = \sum_i p_i H_k(p_{r,j} \mid r = i).$$

We have

$$\begin{aligned} H_k(p_{i,j} \mid p_i) &= -\sum_i p_i \sum_j \frac{p_{i,j}}{p_i} \log_k \frac{p_{i,j}}{p_i} \\ &= -\sum_{i,j} p_{i,j} \log_k p_{i,j} + \sum_i p_i \log p_i \\ &= H_k(p_{i,j}) - H_k(p_i). \end{aligned}$$

Thus,

$$H_k(p_{i,j}) \geq H_k(p_i).$$

Equality obtains iff $p_{i,j}/p_i \in \{0,1\}$; i.e., $p_{i,j}$ is equal either to zero or to p_i, for any i and j (the marginal distribution determines the complete distribution).

Let $p_1, \ldots, p_n$ be a probability distribution. Assign to each i, $1 \leq i \leq n$, a codeword w_i over a k-ary alphabet, so that no codeword is a prefix of another (the code is *instantaneous decodable*). Let

$$L = \sum_{i=1}^{n} p_i |w_i|$$

be the average code length. Then, the Shannon Coding Theorem [10] asserts that the entropy of the distribution is a lower bound on the average code length:

$$L \geq H_k(p_i).$$

Equality obtains iff

$$|w_i| = -\log_k p_i$$

for each codeword w_i.

We first derive bounds for the mean path length in the network, in terms of the entropy of the traffic distribution π.

Let

$$\pi_i = \sum_j \pi_{i,j}$$

be the marginal distribution of $\pi_{i,j}$; π_i is the relative frequency of messages generated at input i. We have

Theorem 3.3. *For any route distribution ϱ consistent with π, $d_\varrho \geq H_k(\pi_{i,j} \mid \pi_i)$. Equality obtains only when the traffic from i to j is routed through a unique path of length $-\log_k(\pi_{i,j}/\pi_i)$, for each input-output pair (i,j).*

Proof. Let $\varrho_i(p)$ be the probability that a message issued by input i uses the path p:

$$\varrho_i(p) = \varrho(p)/\pi_i.$$

Let L_i be the average length of a path outgoing input i:

$$L_i = \sum_p \varrho_i(p) \mid p \mid .$$

A path of length l can be encoded, given its origin, by a word of length l over a k-ary alphabet: The word specifies the outgoing link used at each switch on the path (the word is the header the message would carry to specify its route). Since all outputs are sinks, the path to an output does not traverse another output. Thus, no codeword is a prefix of another. Using Shannon's coding theorem, we obtain

$$L_i \geq H_k(\varrho_i(p)).$$

Equality obtains iff

$$|p| = -\log_k \varrho_i(p)$$

for each path p. For each fixed input i, we have

$$H_k(\varrho_i(p)) \geq H_k(\pi_{r,j} \mid r = i).$$

Equality obtains iff all traffic from input i to output j is concentrated on a unique path $p \in \Pi_{i,j}$, for any output j. Summing over all i, we obtain

$$\sum_i \pi_i L_i \geq \sum_i \pi_i H_k(\pi_{r,j} | r = i) = H_k(\pi_{i,j} \mid \pi_i).$$

But

$$\begin{aligned} \sum_i \pi_i L_i &= \sum_i \pi_i \sum_j \sum_{p \in \Pi_{i,j}} \frac{\varrho(p)}{\pi_i} |p| \\ &= \sum_{i,j} \sum_{p \in \Pi_{i,j}} \varrho(p) |p| \\ &= d_\varrho. \end{aligned}$$

Thus,

$$d_\varrho \geq H_k(\pi_{i,j} \mid \pi_i).$$

Equality obtains iff, for each input-output pair (i,j), a unique path $p \in \Pi_{i,j}$ is used to route traffic from i to j, and that path has length

$$|p| = -\log_k \varrho_i(p) = -\log_k(\pi_{i,j}/\pi_i). \qquad \square$$

We can now plug this estimate of d_ϱ into the bound of Lemma 3.1.

Corollary 3.4. *Let C be the cost and B be the bandwidth of a network G, for traffic distribution π. Then*

$$C \geq B \cdot H_k(\pi_{i,j} \mid \pi_i).$$

Equality obtains if and only if there is a route distribution ϱ consistent with π such that

- *all traffic from i to j uses a unique path of length $-\log_k(\pi_{i,j}/\pi_i)$, for each input-output pair (i,j), and*
- *the relative load $\omega_\varrho(e)$ is equal on each link e.*

We now specialize the last result to networks where the traffic of each input is equally distributed over all outputs, but different inputs may generate different amounts of traffic. We thus have

$$\pi_{i,j} = \pi_i/N.$$

We call such a traffic distribution *semisymmetric*. For example, if inputs are processors and outputs are memory modules, then a semisymmetric distribution assumes that each processor is equally likely to access any memory module, but that distinct processors may generate different amounts of memory traffic.

Corollary 3.5. *Let C be the cost and B be the bandwidth of a network G, for the semisymmetric traffic distribution. Then*

$$C \geq B \cdot \log_k N.$$

Equality obtains iff there exists a route distribution consistent with semisymmetric traffic distribution such that traffic between each input-output pair uses a unique path of length $\log_k N$, and each link has equal relative load.

Proof. We have

$$H_k(\pi_{i,j} \mid \pi_i) = -\sum_i \pi_i \sum_{j=1}^{N} \frac{\pi_{i,j}}{\pi_i} \log_k\left(\frac{\pi_{i,j}}{\pi_i}\right)$$

$$= -\sum_i \pi_i \sum_{j=1}^{N} \frac{1}{N} \log_k\left(\frac{1}{N}\right) = \log_k N. \qquad \square$$

4. Banyan networks and symmetric traffic

4.1. Rectangular banyan networks

A network with a unique path from each input to each output is called a ***banyan network***. A network is *layered* if the underlying graph is acyclic, and all input-output paths have the same length. The nodes in a layered network can be partitioned in layers, with links connecting nodes from one layer to nodes at the next layer. A *degree-k rectangular banyan network* is a layered banyan network where all nodes (except for inputs) have indegree exactly k and all nodes (except for outputs) have outdegree exactly k. A rectangular banyan network that has $M=k^r$ input nodes and $N=k^r$ output nodes is said to have *order r*. It has $(r-1)k^r$ internal nodes. There is a unique path of length r from each input to each output. Figure 1 shows a rectangular banyan network of degree $k=2$ and order $r=2$. It is not unique: nonisomorphic banyan networks of the same degree and order exist (see, e.g. [5]). A rectangular banyan network of order r has cost rk^{r+1}.

The following theorem shows that, with respect to bandwidth, rectangular banyan networks are optimal for symmetric traffic.

Theorem 4.1. *Let G be a network with $M=k^r$ inputs, $N=k^r$ outputs, and outdegree bounded by k. Then the following are equivalent*:

(1) *G is a rectangular banyan network.*

(2) *G has (maximal) bandwidth k^{r+1} for symmetric traffic, and cost rk^{r+1}.*

(3) *G has the best possible bandwidth/cost ratio for symmetric traffic.*

Proof. Note that, for any network G satisfying the hypotheses of this theorem, we have

(a) $C \geq Br$ (Corollary 3.5), and

(b) $B \leq k^{r+1}$ (Lemma 3.2).

Let G be a rectangular banyan network. The number of links of G is equal to

$$C = rk^{r+1}.$$

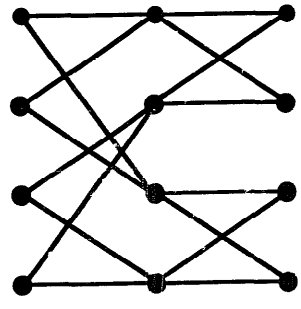

Fig. 1. Rectangular banyan network.

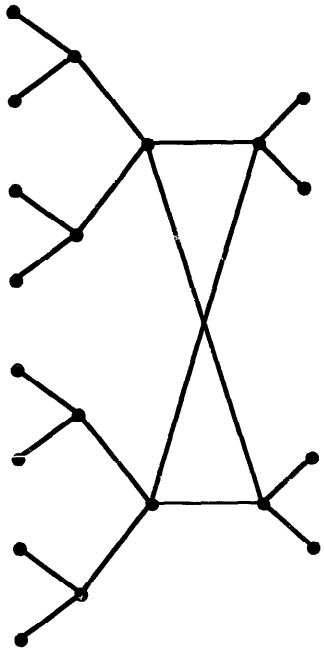

Fig. 2. (8,4)-contracting banyan network.

Each link occurs on exactly k^{r-1} paths. Since there are k^r inputs and k^r outputs, the relative amount of traffic on each path is k^{-2r}. So, the relative load on each edge is $k^{r-1} \cdot k^{-2r} = k^{-(r+1)}$, and the bandwidth is $B = k^{r+1}$. Thus (1) $\Rightarrow$ (2).

By (a), the best possible bandwidth/cost ratio is r. Thus, (2) $\Rightarrow$ (3).

The best possible bandwidth/cost ratio for G is r. Assume that $B = C/r$ for symmetric traffic. Then, by Corollary 3.5, traffic from each input to each output uses a unique path, of length r, and each link has the same relative load. Consider the subgraph containing all paths connecting a fixed input to the k^r outputs. The outdegree of each node in this subgraph is $\leq k$, the input node has indegree 0, the output nodes have outdegree 0, and the distance from the input node to each output node is $\leq r$. This implies that the subgraph is a complete k-ary tree of depth r, and the input node is connected to each output node by a unique path of length r. Thus, G is a layered banyan network, and each node that is not an output has outdegree k. Since each link has the same relative load, a switch has the same number of incoming links as of outgoing links. Thus, each node that is not an input has indegree k. It follows that G is a rectangular banyan network of order r and degree k, and (3) $\Rightarrow$ (1). □

4.2. Contracting banyan networks

Networks of cost (and bandwidth) smaller than rectangular banyan networks can be built out of contracting layers of $k \times 1$ nodes, followed by layers of $k \times k$ nodes,

followed by layers of $1 \times k$ nodes. Formally, an *(M, N)-contracting banyan network of degree k and order r* has $M = k^m \geq k^r$ inputs and $N = k^n \geq k^r$ outputs, and contains a rectangular banyan network of order r. M/k^r inputs are connected to each source of the rectangular banyan network by a balanced binary tree; N/k^r of the outputs are similarly connected to each sink.[1] This network has bandwidth $B = k^{r+1}$ for symmetric traffic, and cost

$$\frac{k}{k-1}(k^m + k^n) + \left(r - \frac{2}{k-1}\right) k^{r+1} = \frac{k}{k-1}(M+N) + B\left(\log_k B - \frac{k+1}{k-1}\right)$$
$$= \Theta(M + N + B \log_k B).$$

Figure 2 shows an (8,4)-contracting banyan network of degree $k = 2$ and order $r = 1$.

We conjecture that a contracting banyan network provides optimal bandwidth for its cost:

Conjecture 1. Let G be an (M, N)-contracting banyan network. There is no acyclic network with k^m inputs, k^n outputs, and outdegree bounded by k that has both lower cost and higher or equal bandwidth for symmetric traffic than G.

While we have not been able to prove this conjecture, we do have two partial results:

(1) The conjecture is true, up to a constant factor.

(2) The conjecture is true for layered networks.

The first claim is a special case of Theorem 5.1 from the next subsection. The proof of the second claim is given in Appendix B.

5. Nonsymmetric traffic

In the previous two subsections, we proved tight bounds on network cost for symmetric traffic. Here we will derive the cost, up to a constant factor, of an optimal network for semisymmetric traffic for any bandwidth. We then present upper and lower bounds for completely general traffic distributions.

Let $\pi_{i,j}$ be a traffic distribution for M inputs and N outputs; let $\pi_i = \sum_j \pi_{i,j}$, and let $\pi^j = \sum_i \pi_{i,j}$. In a network with outdegree bounded by k, the total amount of traffic outgoing input i is at most k messages per time unit; thus, we must have

$$B \leq \min_i k/\pi_i. \tag{2}$$

Similarly, the indegree constraints at the output nodes imply that

$$B \leq \min_j k/\pi^j. \tag{3}$$

[1] Variants of contracting and "expanding" banyan networks have been defined and studied [1,6]. We have taken the liberty of modifying their definitions to fit our context.

Recall that in semisymmetric traffic the input nodes have different traffic intensities, but traffic from each input node is equally distributed to all output nodes: $\pi_{i,j}=\pi_i/N$, and $\pi^j=1/N$.

Theorem 5.1. *Let π be a semisymmetric traffic distribution on M inputs and N outputs. Let B fulfill inequalities* (2), (3). *Then the minimal cost of a network with indegree and outdegree bounded by k that supports π with bandwidth B is* $\Theta(M+N+B\log_k B)$.

Proof. To prove the lower bound, consider a network satisfying the hypotheses. By Corollary 3.5, it has cost $\Omega(B\log_k N)$. Since the network is connected, it has at least $\Omega(M+N)$ links, Thus, the cost is $\Omega(M+N+B\log_k N)=\Omega(M+N+B\log_k B)$.

To prove the upper bound, we construct a network. Assume, w.l.o.g., that $B=k^b$. Let $S_0=\{i\colon \pi_i\le 1/B\}$, and $S_1=\{i\colon \pi_i>1/B\}$; let $w_0=\sum_{i\in S_0}\pi_i$ and $w_1=\sum_{i\in S_1}\pi_i$. Using a first fit decreasing bin packing algorithm, we can pack the items π_i, for $i\in S_0$, into bins of size $1/B$, so that each bin, with the possible exception of the last, is at least half full. The number of bins used is $\le\lceil 2w_0B\rceil$. It follows that the inputs in the set S_0 can be associated with the leaves of $\le\lceil 2w_0B/k\rceil$ k-ary trees, so that the following holds: (1) The sum of the weights in each tree is $\le k/B$, and (2) the sum of the weights in each proper subtree is $\le 1/B$. Each of the inputs in the set S_1 is associated with a trivial, one-node tree; since $\pi_i\le k/B$, for each i, property (1) holds for these trees, too. The total number of trees used is

$$\begin{aligned}\le\lceil 2w_0B/k\rceil+|S_1| &\le\lceil 2w_0B/k\rceil+\lfloor w_1B\rfloor\\ &\le\lceil w_0B\rceil+\lfloor w_1B\rfloor\\ &=B.\end{aligned}$$

The total number of edges in all these trees is $O(M-B)$.

Construct a (B,N)-contracting banyan network G of order b. The network G has bandwidth kB for symmetric traffic, and hast cost $O(N+B\log_k B)$. Extend this network into a contracting (M,N) banyan network G' by connecting each input tree to one of the inputs of G (identifying the root of the tree to an input of G). The resulting network has cost $O(M+N+B\log_k B)$. The traffic from each input of G' is first routed to the root u of the associated tree, next routed in G as traffic from u would be routed. Condition (2) implies that the relative load on each link of an input tree does not exceed $1/B$. Condition (1) implies that each input node u of G receives a fraction of at most k/B of the total network traffic (from inputs in the tree rooted at u). Furthermore, the traffic from u is evenly distributed to all N outputs. It follows that the relative load of each link in G is no more than would occur under symmetric traffic, which is $1/B$. This implies that the bandwidth is at least B. □

Using a construction similar to the randomized routing of Valiant and Brebner [11], we can extend the upper bound result to arbitrary traffic distribution.

A symmetric argument implies that the last theorem is valid for a "reverse semisymmetric" traffic distribution π, where $\pi_{i,j}=\pi_j/M$ (traffic to each output is equally likely to arrive from each input, but distinct outputs may have different amounts of traffic).

Theorem 5.2. *Let π be a traffic distribution on M inputs and N outputs, and let B fulfill inequalities* (2), (3). *Then there exists an M-input, N-output network with indegree and outdegree bounded by k, bandwidth B, and cost $C=O(M+N+B\log_k B)$.*

Proof. The network consists of two halves; the first half has M inputs and $M+N$ outputs; the second half has $M+N$ inputs and N outputs; outputs of the first half are identified with inputs of the second half. A message is routed to a randomly chosen output of the first half, and then routed in the second half to its output. The traffic in the first half is semisymmetric, with distribution $\pi^1_{i,j}=\pi_i/(M+N)$; the traffic in the second half is reserve semisymmetric, with distribution $\pi^2_{i,j}=\pi^j/(M+N)$. We obtain, by the previous theorem and the following remark, that each half can support bandwidth B at cost $O(M+N+B\log_k B)$. □

There can be a gap of up to $\Theta(\log_k B)$ between the lower bound of Corollary 3.4 and the upper bound of the last theorem. Consider, for example, a traffic distribution for $M=N$ inputs and outputs where $\pi_{i,j}=\delta_{i,j}$; each input communicates with a unique output. Clearly, a bandwidth of $B=M$ can be supported in this communication pattern at cost M, by directly connecting each input i to output i. The "mixing" that occurs in the construction of Valiant and Brebner transforms any traffic distribution into a distribution with maximal entropy $\log_k(\max(M,N))$.

However, there is a subtler reason for our failure to achieve tight bounds. It turns out that entropy does not provide a full characterization of the "complexity" of a traffic distribution. Consider, for example, the family $\mathcal{F}$ of traffic distributions $\pi_{i,j}$ on $M=N=k^m$ inputs and outputs with the property that $\pi_{i,j}$ is either equal to zero or to $1/(M\log_k M)$: Each input is equally likely to send messages to a set of $\log_k M$ outputs and, likewise, each output is equally likely to receive a message from a set of $\log_k M$ inputs. All distributions in the family $\mathcal{F}$ have the same entropy $\log_k(M\log_k M)$; the conditional entropy is also the same for all distributions in $\mathcal{F}$, and is equal to $\log_k \log_k M$. Nevertheless, there is a distribution in $\mathcal{F}$ such that the cost of the optimal network is within a $\Theta(\log_k \log_k M)$ factor of the $\Theta(M\log_k M)$ upper bound of Theorem 5.1, and another distribution such that the cost of the optimal network exactly matches the $\Theta(M\log_k \log_k M)$ information theoretic lower bound of Corollary 3.4.

To carry the proof we need to assume that a unique path is used for traffic between an input-output pair, even though Theorem 3.3 does not apply. This assumption is justified by the following theorem.

Theorem 5.3. *Let q be an integer. Let π be a traffic distribution such that $\pi_{i,j} \in \{0, (1/qB)\}$, for any input-output pair i,j. Let G be a network that has bandwidth B for traffic distribution π. Then bandwidth $B_\varrho = B$ can be achieved by a route distribution ϱ compatible with π that routes the traffic between any input-output pair through a unique path.*

The proof of Theorem 5.3 is given in Section 7.

Theorem 5.4. *Let $C(\pi)$ be the minimal cost of a network with $M = N = k^m = k^{k^r}$ inputs and outputs and indegree and outdegree bounded by k that achieves bandwidth $B = kM$ for traffic distribution π. Let $\mathcal{F}$ be the set of traffic distributions, where $\pi_{i,j} \in \{0, 1/(mM)\}$. There are two distributions π^1, $\pi^2 \in \mathcal{F}$ such that $C(\pi^1) = \Omega(M \log_k M / \log_k \log_k M)$ but $C(\pi^2) = O(M \log_k \log_k M)$.*

Proof. For the lower bound, we define a traffic distribution $\pi^2 \in \mathcal{F}$ and construct a network that can support bandwidth kM with "small" cost. Partition the set $\{1, \ldots, M\}$ into M/m subsets $S_1, \ldots, S_{M/m}$, each containing m indices. Consider the traffic distribution where $\pi_{i,j} = 1/(mM)$, if i and j belong to the same set S_k, and $\pi_{i,j} = 0$, otherwise. A bandwidth of kM for this traffic distribution is obtained by a network that consists of M/m disjoint rectangular banyan networks of order r. The total cost of such network is $O((M/m)(mr)) = O(M \log_k \log_k M)$.

For the upper bound, we give a counting argument that shows that some distribution $\pi^1 \in \mathcal{F}$ requires "large" cost C to support bandwidth kM. Let G be a network with cost C and bandwidth kM, for a traffic distribution $\pi \in \mathcal{F}$.

By Theorem 5.3, since the bandwidth $B = kM$ divides mM, it can be achieved by a route distribution that allocates a unique path to each input-output pair. Each such path is used with probability $1/(mM)$. Since the relative load on each link is at most $1/(kM)$, it follows that a link may occur on at most m/k such paths. Let $\hat{G}$ be the graph obtained from G by replicating each link m/k times. The last argument implies that $\hat{G}$ contains link disjoint paths that connect each input-output pair (i,j) such that $\pi_{i,j} > 0$.

Consider $\hat{G}$ as a circuit switching network: Each node has at most m incoming links and m outgoing links; the node can connect each incoming link to an outgoing link in an arbitrary permutation. A set of link disjoint paths correspond to a setting of these switches.

Since each switch of $\hat{G}$ has at most m inputs, a switch with j outputs has $< m^j$ settings. The total number of settings of all switches is bounded by

$$\prod_{u \in \hat{G}} m^{O(u)} = m^{Cm/k} = (\log_k M)^{O(C \log_k M)}.$$

This is an upper bound on the number of distinct traffic distributions $\pi \in \mathcal{F}$ that can be supported by G, with bandwidth kM.

Assume that all traffic distributions in $\mathcal{F}$ can be supported by networks of cost $\leq C$,

with bandwidtk kM. The number of distinct networks with $\leq C$ links, and indegree $\leq k$, is $< C^{kC}$; the number of distributions in $\mathcal{F}$ is $M^{\Omega(M \log_k M)}$. We obtain the inequality

$$C^{kC}(\log_k M)^{O(C \log_k M)} \geq M^{\Omega(M \log_k M)}.$$

This implies that

$$C(\log_k C + \log_k M \log_k \log_k M) = \Omega(M \log_k^2 M),$$

so that

$$C = \Omega(M \log_k M / \log_k \log_k M). \qquad \square$$

6. Network delay

We have so far concentrated on finding minimum cost networks that achieve a particular bandwidth. Another important measure of network performance is message *delay*, the average time required for a message to reach its destination. We desire to attach to each network topology, and each traffic distribution a delay measure. This measure is motivated by the behavior of the M/M/1 system, defined in Appendix A: Messages from i to j are generated by a Poisson process with parameter $\tau\pi_{i,j}$, link transfer times are exponentially distributed i.i.d. random variables with parameter one, service discipline in FCFS, and routing is nonadaptive.

Let ϱ be a routing distribution. Denote by $\tau(e)$, the traffic intensity on link e:

$$\tau(e) = \tau\omega_\varrho(e).$$

The *delay* for a message on a link $e = uv$ is the time from the moment the message arrives to u until it arrives to v; this is the sum of its waiting time at u, and its transfer time on link e. The average delay of a message on link e is equal to

$$\frac{1}{1-\tau(e)}.$$

The average delay for a message using path p is

$$\sum_{e \in p} \frac{1}{1-\tau(e)}$$

[4, §3.1, ex. 4]. It follows that the average time for a message to reach its destination is

$$D_\varrho(\tau) = \sum_p \varrho(p) \sum_{e \in p} \frac{1}{1-\tau(e)} = \sum_e \frac{\omega_\varrho(e)}{1-\tau\omega_\varrho(e)}.$$

We take D_ϱ to be the *delay measure function* for network G, with route distribution ϱ.

The delay D_ϱ equals to the sum of the average delay on each link, where each link is weighted by its relative load. The delay goes to infinity when the traffic intensity τ approaches the maximal intensity $B = \min 1/\omega_\varrho(e)$.

6.1. Lower bounds

We can obtain a lower bound on delay in terms of the mean path length d_ϱ and the cost C:

Lemma 6.1.

$$D_\varrho(\tau) \geq \frac{d_\varrho}{1 - \tau d_\varrho / C}.$$

Equality is achieved iff all links have the same load.

Proof. We have

$$d_\varrho = \sum_e \omega_\varrho(e)$$

(see Lemma 3.1). We minimize, for fixed τ,

$$D_\varrho(\tau) = \sum_e \frac{\omega_\varrho(e)}{1 - \tau\omega_\varrho(e)}$$

under the constraint

$$\sum_e \omega_e = d_\varrho.$$

The minimum occurs when all ω_e are equal. The result is obtained by substitution. □

Note the similarity between the bound in Lemma 6.1 for average delay, and the formula for system time in M/M/1 queueing systems: d_ϱ is the average service time for a message in the network, and d_ϱ/C is the utilization factor for the network.

Corollary 6.2. *Let G be a network with outdegree bounded by k. Then*

$$D_\varrho \geq \frac{H_k(\pi_{i,j} \mid \pi_i)}{1 - \tau H_k(\pi_{i,j} \mid \pi_i)/C}.$$

Equality obtains iff all links have equal load and a unique path of length $-\log_k(\pi_{i,j}/\pi_i)$ is used to route traffic from input i to output j, for all entry-exit pairs.

Proof. By Theorem 3.3, we can substitute entropy for mean path length. □

Corollary 6.3. *Let G be a network with M inputs, N outputs, and outdegree bounded by k. Then, for ϱ consistent with semisymmetric traffic,*

$$D_\varrho \geq \frac{\log_k N}{1 - \tau \log_k N/C}.$$

Equality obtains iff all links have equal load and a unique path of length $\log_k N$ *is used to route traffic from each input to each output.*

The condition for equality in the last corollary is the same condition that implies that a network has optimal cost/bandwidth ratio (Theorem 4.1). We obtain:

Theorem 6.4. *Let G be a network with* $N=M=k^m$ *inputs and outputs, and in-degree and outdegree bounded by* k. *Then, for* ϱ *consistent with symmetric traffic,*

$$D_\varrho \geq \frac{m}{1-\tau m/C}.$$

Equality obtains (for any feasible value of τ*) iff G is a rectangular banyan network.*

Thus, rectangular banyan networks uniquely achieve an optimal delay-to-cost relation: Any other network, with the same cost, has worse delay, at any traffic intensity.

7. Multicommodity flows

We can formalize the problem of finding a route distribution that maximizes bandwidth as a *constrained multicommodity flow problem*. For a similar approach see [9]. This formalization will allow us to use integer solution theorems from linear programming in order to restrict the type of route distributions that need to be considered.

A multicommodity flow problem is defined by the following:

- A directed flow network $G=\langle V,E\rangle$.
- An assignment of a nonnegative capacity c_e to each link $e\in E$.
- A set of commodities $1,\ldots,h$.
- A source s_l and a sink t_l for each commodity l.

A *flow* is defined by a set of variables x_{el}, where x_{el} is the flow of commodity l through link e. A flow is feasible if

- $x_{el}\geq 0$, for each link e and each commodity l;
- $\sum_l x_{el}\leq c_e$, for each link e;
- $\sum_{e\in I(u)} x_{el}=\sum_{e\in O(u)} x_{el}$, for each node u and each commodity l, $u\neq s_l,t_l$.

The *value* of the flow in commodity l is equal to

$$v_l=\sum_{e\in I(t_l)} x_{el}=\sum_{e\in O(s_l)} x_{el}.$$

The total value of the flow, which we attempt to maximize, is

$$v=\sum_l v_l.$$

We impose an extra constraint of the ratios of the flow values:

- $v_i = \lambda \phi_i$, $\forall i$ $(1 \le i \le h)$, where $\phi = \langle \phi_1, \ldots, \phi_h \rangle$ is a nonnegative vector.

The total flow value v is maximum when λ is maximum.

Let G be an interconnection network, and π be a traffic distribution in G. Assign to each link of G a capacity of one. Consider the following constrained multicommodity flow problem for this network: there is a distinct commodity for each input-output pair i, j, with source i and sink j. The ratios between the flow values in each commodity are defined by the vector $\pi_{i,j}$.

Theorem 7.1. *The bandwidth B of network G for traffic distribution π is equal to the maximum total flow value for the constrained multicommodity flow problem defined above. Moreover, if x_{eij} is an optimal solution to the constrained flow problem, then there is a corresponding route distribution ϱ that achieves optimal bandwidth $B_\varrho = B$, such that x_{eij} is the amount of traffic from input i to output j routed through link e. For each input-output pair (i,j) if $p \in \Pi_{i,j}$, then $B\varrho(p)$ is a multiple of the g.c.d. of $\{x_{eij}: e \in G\}$.*

Proof. Let ϱ be a route distribution that achieves optimal bandwidth $B_\varrho = B$. Define x_{eij} to be the relative load on link e due to traffic from input i to output j, i.e.,

$$x_{eij} = \sum_{p \in \Pi_{i,j}, e \in p} \varrho(p).$$

Then it is easy to see that x_{eij} is a feasible solution to the constrained multicommodity flow problem, with total flow value B.

Conversely, assume that x_{eij} is an optimal solution for the constrained multicommodity flow problem, with value v_{ij} in commodity (i,j), and total value $v = \sum v_{ij}$. Then v_{ij} is a solution to the following (single commodity) flow problem.

Maximize

$$a = \sum_{p \in \Pi_{i,j}} \xi_p,$$

subject to

$$\sum_{e \in p} \xi_p \le x_{eij}, \quad \xi_p \ge 0.$$

By the integral flow theorem, we can assume that each variable ξ_p is a multiple of the g.c.d. of $\{x_{eij}: e \in G\}$. Define $\varrho(p) = \xi_p / v$. Then $\varrho(p)$ is a route distribution that is consistent with π, and achieves bandwidth $B_\varrho = v$. □

In the single commodity case a flow problem with integer coefficients has an integer optimal solution [7, Theorem 2.1]. This is not true for multicommodity flow problems: even if all capacities are integer, an optimal flow may have noninteger components. However, if there is a feasible solution where all (source-to-sink) commodity flow values are integer, then one can achieve the same commodity flow values with a solution where flows of each commodity on each link are integer.

Theorem 7.2. *Let $\langle v_1, \ldots, v_h \rangle$ be the values of a feasible multicommodity flow for the network G, with capacities c_e, sources $s_1, \ldots, s_h$, and sinks $t_1, \ldots, t_h$. Assume that the v_i and c_e are all integer. Then there exists a feasible multicommodity flow in this network with values $\langle v_1, \ldots, v_h \rangle$ where all flows are integer.*

Proof. Let

$$v_{ul} = \begin{cases} -v_l, & \text{if } u = s_l, \\ v_l, & \text{if } u = t_l, \\ 0, & \text{otherwise.} \end{cases}$$

Let $\boldsymbol{G}$ be the incidence matrix of the graph G:

$$\boldsymbol{G}_{ue} = \begin{cases} 1, & \text{if } e \in I(u), \\ -1, & \text{if } e \in O(u), \\ 0, & \text{otherwise.} \end{cases}$$

Let s_e be slack variables. Then x_{el} is a feasible flow with values $\langle v_1, \ldots, v_h \rangle$ iff it is a solution to the linear system

- $\sum_l x_{el} + s_e = c_e$, for each e,
- $\sum_e \boldsymbol{G}_{ue} x_{el} = v_{ul}$, for each u and l,
- $x_{el}, s_e \geq 0$.

The matrix of the system of equations has the form

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G} & \cdots & \boldsymbol{0} & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{I} & \boldsymbol{I} & \cdots & \boldsymbol{I} & \boldsymbol{I} \end{bmatrix}.$$

The matrix $\boldsymbol{G}$, which is the incidence matrix of a directed graph, is totally unimodular: each submatrix has determinant $+1$, -1, or 0 [3]. The determinant of a nonsingular submatrix of $\boldsymbol{A}$ is the product of determinants of submatrices of $\boldsymbol{G}$; thus $\boldsymbol{A}$ is totally unimodular. It follows that the linear system has an integer valued solution [3]. □

Corollary 7.3. *Let G be a network with bandwidth B for traffic distribution π. Let β be the g.c.d. of 1 and $\{B\pi_{i,j}\}$. Then there exists a route distribution ϱ consistent with π that achieves bandwidth $B_\varrho = B$, such that $B\varrho(p)$ is a multiple of β, for any path p.*

Proof. Consider the multicommodity flow problem associated with the network G and the distribution π. By Theorem 7.1 this problem has an optimal solution with

value B, the network bandwidth. By Theorem 7.2 there is an optimal solution such that all flows x_{eij} are a multiple of the g.c.d. of the source-to-sink flow values, which are $B\pi_{i,j}$, and of the link capacities, which are all 1. It follows that all flows x_{eij} are a multiple of β. By Theorem 7.1, the bandwidth B can be realized by a route distribution ϱ, such that, for each path p, $B\varrho(p)$ is a multiple of the g.c.d. of $\{x_{eij}\}$, which is a multiple of β. □

We can now prove Theorem 5.3, which we restate for convenience.

Theorem 5.3. *Let q be an integer. Let π be a traffic distribution such that $\pi_{i,j} \in \{0, 1/qB\}$, for any input-output pair i, j. Let G be a network that has bandwidth B for traffic distribution π. Then bandwidth $B_\varrho = B$ can be achieved by a route distribution ϱ compatible with π that routes the traffic between any input-output pair through a unique path.*

Proof. The g.c.d. of 1 and $B\pi_{i,j}$ is equal to $1/q$. Thus, by the previous corollary, a bandwidth of B is achieved by a route distribution ϱ where all path probabilities are multiples of $1/qB$. If $\pi_{i,j} \neq 0$, then $\sum_{p \in \Pi_{i,j}} \varrho(p) = \pi_{i,j} = 1/qB$, and $\varrho(p)$ are all multiples of $1/qB$. It follows that $\varrho(p) > 0$ for a unique path $p \in \Pi_{i,j}$. □

Consider, for example, the symmetric traffic distribution $\pi_{i,j} = 1/MN$. Then any bandwidth $B = MN/r$, where r is an integer, can be achieved when using a unique path to connect each input to each output. The use of a single path for traffic between each pair of points simplifies routing. The last theorem implies that no loss of performance is entailed, as far as bandwidth is concerned. Using multiple paths may still improve other performance parameters, such as delay. Also, the last theorem does not imply that in an optimal network topology there is a unique path between each input and each output; it only implies that a unique path will be used. However, we conjecture that in an optimal network topology there is in fact a unique path between each input and each output.

8. Conclusion

We have presented in this paper a framework whereby one can study the relationship between a network topology and its bandwidth for a particular traffic distribution. Many problems are left open.

We still do not have a constructive way of building near optimal topologies for an arbitrary given traffic distribution and given bounds on link count and node degrees.

The entropy function, while giving useful information on a traffic distribution, does not fully characterize its "complexity". It would be very interesting to study other complexity functions for distributions.

We gave a proof of optimality for delay for rectangular banyan networks. We conjecture that contracting banyan networks also have optimal delays. While increasing cost above that of a rectangular banyan network cannot increase bandwidth, it can decrease delays. One can consider *expanding banyan networks*. Such network G has $M=k^m$ inputs, $N=k^n$ outputs and contains a rectangular banyan network of order r, with $r \geq m, n$; each input of G is connected by a complete binary tree to k^{r-m} inputs of the rectangular banyan network, and similarly for outputs. We conjecture that such networks have optimal delays.

A similar framework can be used to study other figures of merit for networks, such as fault tolerance.

Finally, we have considered in this paper "open networks" where inputs are disjoint from outputs. A similar theory of "closed networks", where inputs coincide with outputs, should be established.

Acknowledgement

The authors thank Marty Reiman for pointing out the publications of Harrison [2] and Kelly [4], David Matula for communicating to us his results [9], and the anonymous referees for their useful comments.

Appendix A: Bandwidth can be achieved

This appendix shows that the bandwidth B defined in Section 2 is the least upper bound on the traffic intensity that the network can support for a given traffic distribution. Consider first the discrete model: K messages, distributed according to the distribution $\pi_{i,j}$, are to be transferred through the network. Assume that the total number of nodes in the network is n. Let ϱ be a route distribution that is consistent with π. We then have

Theorem A.1. *There is a schedule that routes K messages, according to route distribution ϱ, in time $O(K/B_\varrho + n)$, using constant size queues at switches.*

Proof. Leighton et al. [8] show that there is a schedule of length $O(c+d)$ for any set of paths such that no link occurs in more than c paths, and no path has length more than d. In our case, since only simple paths are used for routing, path length is bounded by n. By definition, the total number of messages sent on any link is bounded by K/B_ϱ. The result follows. □

The definition can be similarly motivated in the continuous model. Model each link as a server. A message is served for one service period by each link on its path; average service time is 1. Assume that the mean number of new messages generated

per time unit is τ. The expected number of arrivals at link e per time unit is $\tau\omega_\varrho(e)$. Thus, $\tau < B_\varrho$ is a *local equilibrium condition*: arrival rate at any server is lower than service rate. For many distributions, this condition is also sufficient for global equilibrium of the network.

A particular, simple probabilistic model where this holds true is the M/M/1 model, defined as follows:

- Messages are generated at input i for output j by a Poisson process with parameter $\tau\pi_{i,j}$. Traffic between distinct input-output pairs it not required to be independent.
- The transfer times of messages at links are i.i.d. random variables with exponential distribution with parameter 1.
- A first come, first serve (FCFS) queueing discipline is used at each link.
- Routing in the network is nonadaptive: A message going from input i to output j is randomly assigned to a path connecting i to j; path p is chosen with probability $\varrho(p)/\sum_{\hat{p}\in\Pi_{i,j}}\varrho(\hat{p})$.

There is a simple product form for the distribution of queue lengths in M/M/1 network defined by these conditions (see [4, Chapter 3]). If the local equilibrium condition holds then the network has a stationary distribution with bounded moments; the distribution of each queue length is geometric.

We also have the following result, which holds for arbitrary distributions. Consider the following G/G/1 model:

- For each pair i, j messages are generated at input i to output j by a process with independent increments, so that the expected number of messages generated per time unit is $\tau\pi_{i,j}$.
- Transfer times at links are i.i.d. random variables with expectation 1 (in particular, service time may be constant 1).
- Routing is as for the M/M/1 model.

Theorem A.2. *There exists a service policy for messages so that the above* G/G/1 *network has a nondefective stationary distribution when it fulfills the local equilibrium condition.*

(A distribution is nondefective if the underlying random variable is almost surely finite.)

Proof. We use a service policy that decouples service on behalf of distinct paths at each link. Whenever a service period ends at link e, a new path p is chosen with probability

$$\varrho(p)/\omega_\varrho(e).$$

A service period is then spent on behalf of that path; if there is a message waiting to be forwarded on that path, then the first such message is handled; otherwise the link is idle for one service period.

Consider now a fixed path $p = e_1, \ldots, e_k$. The interarrival times of messages to the path p are i.i.d. random variables; the arrival rate of messages onto path p is equal to $\tau\varrho(p)$. These messages are served by k successive servers. Queueing policy at each server is first come first serve (FCFS).

Consider the service of one server e_j on behalf of path p. Let x_n be the time between the end of the $(n-1)$th service period and the end of the nth service period that e_j reserves for path p. Then x_n is the sum of v independent service periods, where v is a random variable that has a geometric distribution with parameter $\varrho(p)/\omega_\varrho(e)$. As each service period has expected length 1, the expected value of x_n is

$$\mathrm{E}(x_n) = \mathrm{E}(v) = \omega_\varrho(e)/\varrho(p).$$

If a message arrives at e_j when other messages from the same path are waiting, then its "service time" will be the length of some period x_n; if it arrives when there are no waiting messages from the same path, then its "service time" will be part of such a period (the residual part of the current period x_n); thus, service rate is at least $\mathrm{E}(x_n)^{-1} = \varrho(p)/\omega_\varrho(e)$. The variables x_n are independent, so that the successive service times are dominated by independent random variables. Thus, each path can be viewed as an independent tandem system of servers, with the output from e_j being the input to e_{j+1}. If the local equilibrium condition is satisfied, then the arrival rate to this system is

$$\tau\varrho(p) < B_\varrho \varrho(p) \leq \varrho(p)/\omega_\varrho(e).$$

The arrival rate to the tandem system is smaller than the service rate at each server. It follows that the system has a nondefective stationary distribution (see [2]). □

This service policy can be implemented by a distributed online probabilistic algorithm. However, this policy is not practical as it significantly increases delays and queue lengths.

Appendix B: Optimality of contracting banyan networks

Theorem B.1. *Let G be a layered network with $M = k^m$ inputs, $N = k^n$ outputs, and bandwidth $B = k^b$ for symmetric traffic, such that G has indegree and outdegree bounded by k. Then*

$$C(G) \geq \frac{k}{k-1}(M+N) + B\left(\log_k B - \frac{k+1}{k-1}\right).$$

Proof. Let ϱ be an optimal route distribution for G, that achieves bandwidth $B_\varrho = B$. We denote by $\omega_\varrho(u)$ the *relative load* of node u:

$$\omega_\varrho(u) = \sum_{e \in I(u)} \omega_\varrho(e) = \sum_{e \in O(u)} \omega_\varrho(e).$$

Our proof uses two arguments. The first one is an "entropy" argument, similar to that used in Theorem 4.1. This forces $B \log_k N$ edges. A second argument is a "traffic deficiency" argument: If $B<M$ ($B<N$), then edges close to inputs (outputs) carry only a small amount of traffic; this forces the remaining edge count.

Using the same argument as in the proof of Theorem 3.3, we can derive from Shannon's coding theorem the inequality

$$\sum_p \varrho(p) \sum_{u \in p} \log_k |O(u)| \geq \log_k N.$$

indeed, the left-hand side equals to the average length of a path descriptor, whereas the right-hand side equals to the conditional entropy of the symmetric distribution. Note that

$$\sum_p \varrho(p) \sum_{u \in p} \log_k |O(u)| = \sum_u \omega_\varrho(u) \log_k |O(u)|.$$

Assume that the network G has $r+1$ layers, $L_0, \ldots, L_r$. Let $C_i = \sum_{u \in L_i} |O(u)|$ be the number of links connecting layer i to layer $i+1$. Since fanin is bounded by k, we have $|L_i| \geq M/k^i$; since fanout is bounded by k, we have $|L_{r-i}| \geq N/k^i$.

Let

$$\Delta_i = \frac{C_i}{B} - \sum_{u \in L_i} \omega_\varrho(u) \log_k |O(u)|.$$

We are going to prove that

$$\sum_{i=0}^{r} \Delta_i \geq \frac{k(M+N)}{(k-1)B} + \log_k \frac{B}{N} - \frac{k+1}{k-1}.$$

This will imply the desired result:

$$\begin{aligned} C &= \sum_{i=0}^{r-1} C_i \\ &= B\left(\sum_{i=0}^{r} \Delta_i + \sum_{u \in G} \omega_\varrho(u) \log_k |O(u)| \right) \\ &\geq \frac{k(M+N)}{(k-1)} + B\left(\log_k \frac{B}{N} - \frac{k+1}{k-1} \right) + B \log_k N \\ &= \frac{k(M+N)}{(k-1)} + B\left(\log_k B - \frac{k+1}{k-1} \right). \end{aligned}$$

For each link e we have $\omega_\varrho(e) \leq 1/B$. Thus, $\omega_\varrho(u) \leq |O(u)|/B$, for each node u that is not an output. This implies that, for $i<r$,

$$\sum_{u \in L_i} \omega_\varrho(u) \log_k |O(u)| \leq \frac{1}{B} \sum_{u \in L_i} |O(u)| \log_k |O(u)|.$$

We also have, for each node $u \in L_i$,

$$0 < |O(u)| \leq k,$$

and

$$\sum_{u \in L_i} |O(u)| = C_i.$$

The maximum of the sum

$$\sum_{u \in L_i} |O(u)| \log_k |O(u)|$$

under these two constraints is equal to C_i. Thus,

$$\sum_{u \in L_i} \omega_Q(u) \log_k |O(u)| \leq \frac{C_i}{B}.$$

This implies that, for $0 \leq i < r$,

$$\Delta_i \geq 0. \tag{4}$$

Consider the first $m-b+1$ layers of the network. A node u in layer i is connected to at most k^i inputs. This implies that

$$\omega_Q(u) \leq \frac{k^i}{M}.$$

For integers $r \geq 1$ and $k \geq 2$, we have $\log_k r \leq r-1$. Thus, if $k^i/M \leq 1/B$, then

$$\omega_Q(u) \log_k |O(u)| \leq \frac{k^i}{M} \log_k |O(u)| \leq \frac{|O(u)|-1}{B}.$$

We obtain

$$\begin{aligned} \sum_{u \in L_i} \omega_Q(u) \log_k |O(u)| &\leq \frac{1}{B} \sum_{u \in L_i} (|O(u)|-1) \\ &= \frac{C_i}{B} - \frac{1}{B} \cdot |L_i| \\ &\leq \frac{C_i}{B} - \frac{1}{B} \cdot \frac{M}{k^i}. \end{aligned}$$

Thus, if $i \leq m-b$, then

$$\Delta_i \geq \frac{1}{B} \cdot \frac{M}{k^i}.$$

Summing over the first $m-b+1$ layers, we obtain

$$\begin{aligned} \sum_{i=0}^{m-b} \Delta_i &\geq \frac{M}{B} \cdot \sum_{i=0}^{m-b} \frac{1}{k^i} \\ &= \frac{M}{B} \cdot \frac{1-1/k^{m-b+1}}{1-1/k} \\ &= \frac{kM-B}{(k-1)B}. \end{aligned} \tag{5}$$

Consider now the last $n-b+1$ layers of the network. A node $u \in L_{r-i+1}$ is connected to at most k^{i-1} outputs, so that $\omega_\varrho(u) \leq k^{i-1}/N$. This implies that, if $u \in L_{r-i}$, then

$$\omega_\varrho(u) \leq \frac{k^{i-1}}{N} \cdot |O(u)|.$$

The maximum of

$$\sum_{u \in L_r} |O(u)| \log_k |O(u)|$$

under the constraints $|O(u)| \leq k$ and $\sum |O(u)| = C_{r-i}$, is equal to C_{r-i}. Thus,

$$\begin{aligned} \sum_{u \in L_{r-i}} \omega_\varrho(u) \log_k |O(u)| &\leq \frac{k^{i-1}}{N} \sum_{u \in L_{r-i}} |O(u)| \log_k |O(u)| \\ &\leq \frac{k^{i-1}}{N} C_{r-i}. \end{aligned}$$

It follows that

$$\begin{aligned} \Delta_{r-i} &\geq C_{r-i}\left(\frac{1}{B} - \frac{k^{i-1}}{N}\right) \\ &\geq |L_{r-i+1}|\left(\frac{1}{B} - \frac{k^{i-1}}{N}\right) \\ &\geq \frac{N}{k^{i-1}}\left(\frac{1}{B} - \frac{k^{i-1}}{N}\right) \\ &= \frac{N}{B} \cdot \frac{1}{k^{i-1}} - 1. \end{aligned}$$

Summing over the last $n-b+1$ layers of links, we obtain

$$\begin{aligned} \sum_{i=1}^{n-b+1} \Delta_{r-i} &\geq \frac{N}{B} \sum_{i=1}^{n-b+1} \frac{1}{k^{i-1}} - (n-b+1) \\ &= \frac{N}{B} \cdot \frac{1 - 1/k^{n-b+1}}{1 - 1/k} - (n-b+1) \\ &= \frac{kN-B}{(k-1)B} - \log_k \frac{N}{B} - 1. \end{aligned} \tag{6}$$

Putting together the inequalities (4)-(6), we obtain

$$\begin{aligned} \sum_i \Delta_i &\geq \frac{kM-B}{(k-1)B} + \frac{kN-B}{(k-1)B} - \log_k \frac{N}{B} - 1 \\ &= \frac{k(M+N)}{(k-1)B} - \log_k \frac{N}{B} - \frac{k+1}{k-1}. \end{aligned}$$

□

References

[1] D. DeGroot, Expanding and contracting sw-banyan networks, in: 1983 International Conference on Parallel Processing (1983) 19–24.

[2] J.A. Harrison, The heavy traffic approximation for single server queues in series, J. Appl. Probab. 10 (1973) 13–29.

[3] A.J. Hoffman and J.B. Kruskal, Integral boundary points of convex polyhedra, in: H.W. Kuhn and A.W. Tucker, eds., Linear Inequalities and Related Systems (Princeton Univ. Press, Princeton, NJ, 1956) 223–246.

[4] F.P. Kelly, Reversibility and Stochastic Networks (Wiley, New York, 1979).

[5] C.P. Kruskal and M. Snir, A unified theory of multistage interconnection network structure, Theoret. Comput. Sci. 48 (1986) 75–94.

[6] M. Kumar and J.R. Jump, Generalized delta networks, in: 1983 International Conference on Parallel Processing (1983) 10–18.

[7] E.L. Lawler, Combinatorial Opzimization: Networks and Matroids (Holt, Rinehart and Winston, New York, 1976).

[8] T. Leighton, B. Maggs and S. Rao, Universal packet routing algorithms, in: Proceedings of the 29th Symposium on Foundations of Computer Science (1988) 256–269.

[9] D.W. Matula, Concurrent flow and concurrent connectivity in graphs, in: Y.A. et al., eds., Graph Theory and its Applications to Algorithms and Computer Science (Wiley, New York, 1985) 543–559.

[10] C. Shannon, A mathematical theory of communication, Bell System J. 28 (1948) 656–715.

[11] L.G. Valiant and G.J. Brebner, Universal schemes for parallel communication, in: Proceedings of the 13th Annual ACM Symposium on Theory of Computing (1981) 263–277.